HP sx2000 チップセット搭載プラットフォームの RAS 機能

White paper

目次

はじめに 3

システム トポロジー 3

インテル Itanium 2 プロセッサ アーキテクチャの RAS 機能 5
- インテル キャッシュ セーフ テクノロジー 5
- プロセッサの動的割り当て解除 (Dynamic Processor Resiliency: DPR) 5

メモリ サブシステムの RAS 機能 6
- ダブル チップ スペアリング (メモリ エラー訂正) 6
- 独自に DIMM を改良 7
- メモリの動的割り当て解除 (Dynamic Memory Resiliency: DMR) 7

I/O サブシステムの RAS 機能 8
- I/O アーキテクチャの概要 8
- マルチパス I/O 8
- 高度な I/O エラー復旧 9
- PCI OL* 10

システム通信リンクの RAS 機能 10
- リンク レベルのリトライとリンクの再学習 10
- 予備のチャネル/予備の回路 11

キャビネット機構 12
- クロスバー バックプレーンの障害復旧力 12
- クロックの冗長化 13
- 電源機構 15
- 冷却機構 16
- 管理プロセッサ 16
- グローバルな共有メモリ 16
- ファームウェアの更新 17

障害管理 18
- 保守性 19
- ハードウェアの予防保守設計 20

まとめ 22

参考情報 23

はじめに

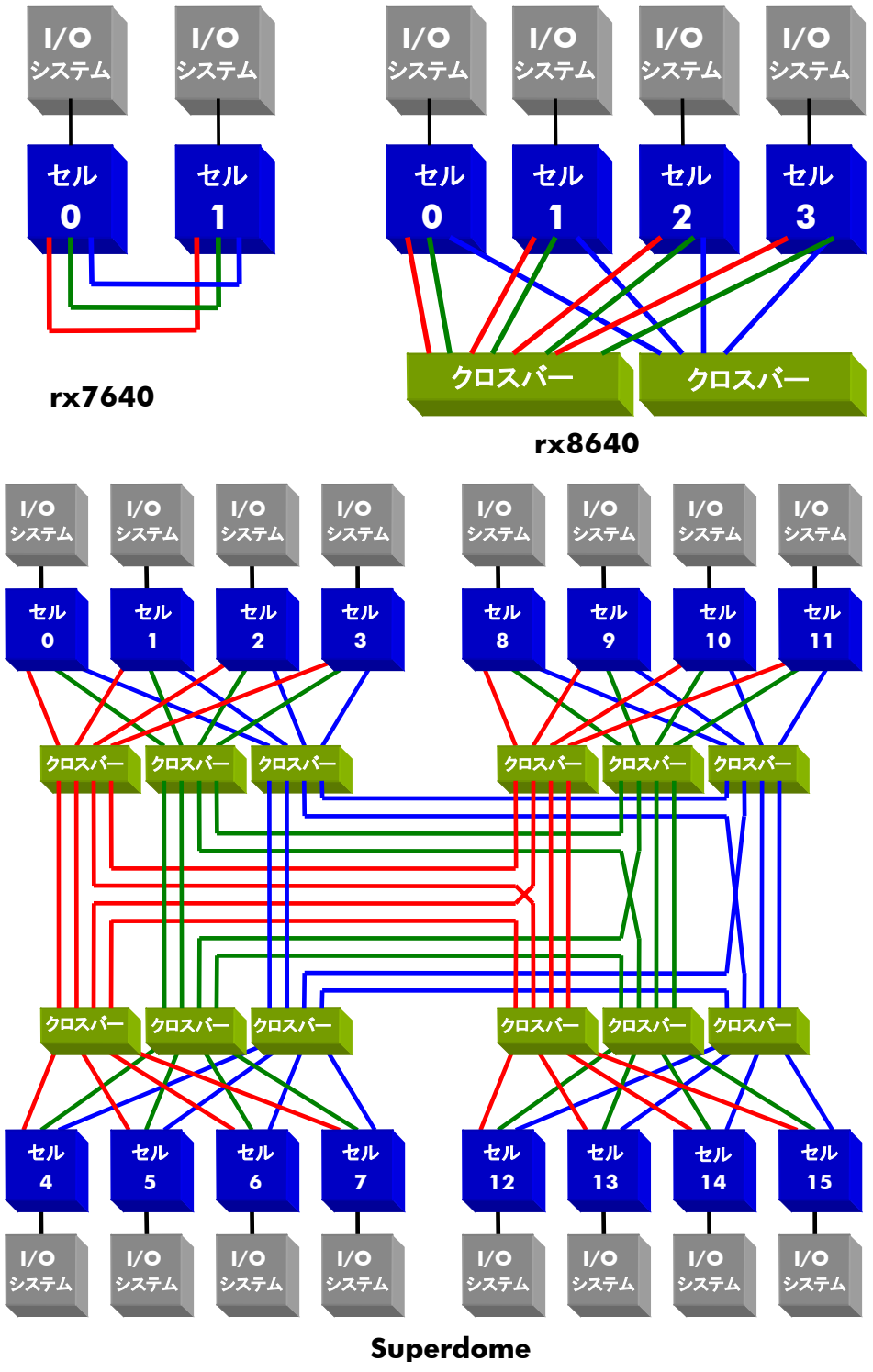
HP の新しい HP sx2000 チップセット搭載の HP Integrity サーバはメインフレーム級のシステムであり、高レベルの可用性を提供します。本書では、高レベルの可用性を実現している具体的な機能について説明します。HP は、システムの各機能領域において、理論的に、または現場での観察を通じて障害モードの分析を行い、従来は HP のフォールト トレランス製品ラインのみに提供していた多くのテクノロジーを第一世代の HP Integrity Superdome の機能に追加することにより、このクラスで最も堅牢なシステムを構築しました。本書では、主に SSHA (Single-System High-Availability: システム単体での高可用性) について説明します。高可用性クラスタリングについては、HP の別のホワイトペーパーおよび技術概要書で説明しています。

システム トポロジー

HP sx2000 チップセットを搭載している HP Integrity サーバは、「セルベース システム」といいます。すなわち、システムはセル ボード (セル) と呼ばれるビルディング ブロックで構成されています。各セルは、プロセッサ、DIMM メモリ、および I/O サブシステムとの接続コンポーネントで構成されます。これらのサーバでは、SMP (Symmetric Multiprocessing: 対称型マルチプロセッシング) アーキテクチャが活用されるほか、セルによりハード パーティション (nPartitions) の構成も可能になっています。ハード パーティションとは、専用のメモリ空間を持つ、互いに分離されたハードウェア リソースのことであり、個々のハード パーティションでは他のハード パーティションから独立してオペレーティング システムのインスタンスを実行できます。ハード パーティションは、単体のサーバ システム内で、1 つ以上のセルから構成されます。1 つのサーバを複数のハード パーティションで構成することにより、サーバ コンソリデーションが可能になります。各パーティションには異なるオペレーティング環境を設定できるので、たとえば 1 つのサーバ上で HP-UX 11i、Microsoft Server 2003、Linux、および OpenVMS を同時に実行することができます。HP-UX を実行するハード パーティションでは、ここからさらに仮想パーティション (vPars) を使用した分割が可能で、サーバのコンソリデーションやバーチャリゼーションに対してきめ細かなパーティションを設定することができます。

HP sx2000 チップセットは、セル コントローラ、メモリ バッファ、クロスバー スイッチ、PCI-X システム バス アダプタ、および PCI-X ホスト ブリッジの 5 種類の VLSI コンポーネントで構成されています。システム バックプレーン上のクロスバー スイッチにより、システム全体のトポロジーが決定されます。図 1 に示すように、HP Integrity rx7640 では、2 つのセルが直接接続されているため、クロスバー スイッチは使用されていません。HP Integrity rx8640 の場合は、2 つのクロスバー スイッチで相互接続されているセルが 4 つあり、3 つのシステム ファブリックで構成されています。HP Integrity Superdome では、2 つのキャビネットのそれぞれに、8 つのセルと 6 つのクロスバー スイッチがあります。このようなクロスバー ベースのトポロジーでは、アップグレードや保守が必要になったハード パーティションやセル (オペレーティング環境でサポートされている場合) を、システムの他のコンポーネントに影響を与えることなく、個別にオフラインにすることができます。これはリング ベースのトポロジーよりも優れています。リング ベースのトポロジーでは、1 つのセルを取り外すとシステム全体の通信が行えないので、システム全体をシャットダウンする必要があるからです。HP sx2000 チップセットのセル ボード アーキテクチャの詳細については、HP sx2000 チップセットの技術概要を参照してください。

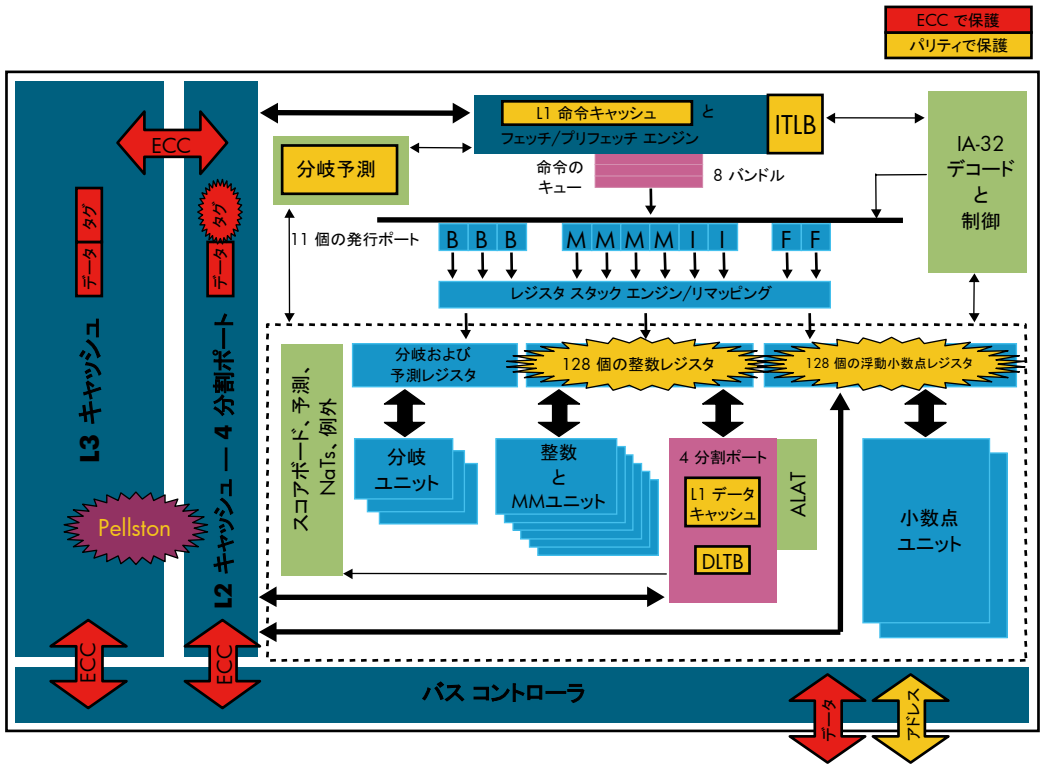
図 1. システム トポロジー



インテル Itanium 2 プロセッサ アーキテクチャの RAS 機能

インテル® Itanium® 2 プロセッサは、エンタープライズ環境でメインフレーム レベルの可用性の実現を目的に設計された、パワフルなエラー訂正スキームと業界最高水準のマシン チェック アーキテクチャを持っています。インテル Itanium 2 プロセッサは、HP sx2000 チップセット搭載システムの高度な RAS 機能を組み合せることにより、従来はハイエンドの専用プラットフォームでしか提供されなかったレベルの可用性をサポートします。

図 2. インテル Itanium 2 プロセッサのアーキテクチャ



インテル Itanium プロセッサ アーキテクチャの堅牢性は、ロス アラモス国立研究所 (LANL) における一連のテストで確認されています。システムは、LANL の中性子ビームを使って、海抜ゼロで観察される粒子束密度の 2 千万倍も濃い環境で試験されました。その結果を一般的なデータセンタの条件に当てはめると、インテル Itanium プロセッサの平均故障時間 (MTTF) は 500 年以上になることが分かりました。また、マシン チェック アーキテクチャによって、致命的なエラーをすべて防止できました。これはローエンドのプロセッサにはない機能です。

インテル キャッシュ セーフ テクノロジー

キャッシュサイズが大きくなれば、キャッシュの故障率が増大します。インテル Itanium プロセッサ シリーズでは、デュアルコア インテル® Itanium® 2 プロセッサ 9000 番台 (開発コード名: Montecito) からインテル キャッシュ セーフ テクノロジー (開発コード名: Pellston) を導入し、この故障率の増大を抑制します。キャッシュ セーフ テクノロジーは、L3 キャッシュに訂正可能なエラーが見つかると、キャッシュ ラインをテストしてエラーを訂正します。キャッシュ ラインに欠陥が見つかった場合、そのラインを無効にします。プロセッサは、本来は無効にしたキャッシュ ラインにマッピングされるはずのデータをフェッチする場合、メモリから直接フェッチします。障害によって影響を受けるのはそのキャッシュ ラインだけなので、性能上の問題は発生しません。キャッシュ セーフ テクノロジーによって、キャッシュ ラインが故障しても、プロセッサがクラッシュすることはありません。

プロセッサの動的割り当て解除 (Dynamic Processor Resiliency: DPR)

DPR はシステムの機能で、訂正可能なエラーの許容数を超えたプロセッサを割り当て対象から自動的に切り離します。訂正されたキャッシュ エラーの数が設定可能な特定のしきい値に到達すると、プロセッサは構成対象から自動

的に切り離されます。DPR は HP-UX では現在サポートされていますが、Windows オペレーティング システムでは限られた範囲でサポートされています。

DPR とキャッシュ セーフ テクノロジーを備えた HP sx2000 チップセット搭載システムには、プロセッサのキャッシュ エラーに対する完全な回復力が備わっているので、システムのダウンタイムが大幅に縮小します。DPR の副作用としては、プロセッサが構成対象から切り離されることによるシステムの処理能力の低下がありますが、この問題は、iCAP を利用してシステムに使用開始状態または停止状態のプロセッサを追加し、システムに予備の処理能力を構成しておくことにより回避できます。

プロセッサ テクノロジーの進歩に伴い、DPR も進化してきたので、新たに復旧可能となったエラーも処理できるようになりました。これも、このような保護機能を持たないローエンドのプロセッサに対するインテル Itanium 2 プロセッサの優位性です。この機能によってプロセッサ エラーのおよそ 70% を解消できます。

メモリ サブシステムの RAS 機能

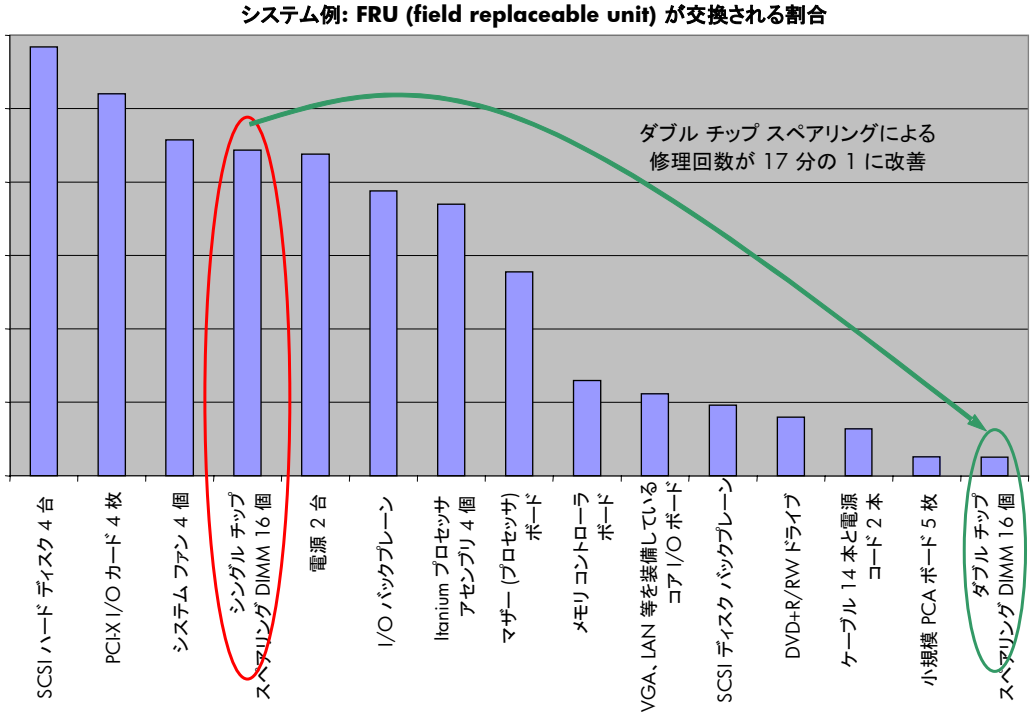
メモリ システムは 288 ビットのメモリ ECC ワードで構成されています。このうち、264 ビットはデータ/タグ ビットとして使用され、24 ビットはそのワード用の ECC チェック ビットとして使用されます。ECC コードは SDRAM チップ スペアリング機能のサポート要件を満たすように設計されています。読み込みサイクルでは、各 x4 SDRAM から 1 つの 144 ビット メモリ チャンクの 4 ビットが、また次の 144 ビット メモリ チャンクの 4 ビットが読み込まれて、ECC コード ワードの 8 つの連続したビットとして扱われます。したがって、1 つの SDRAM が故障すると、その故障は、キャッシュ ライン内の各 4 つの ECC コード ワードで 1 バイトのエラーとして現れます。これらの 4 つのエラーは訂正可能なので、SDRAM の故障は、訂正不可能なエラーの原因とはなりません。1 つの SDRAM の 8 ビットは、すべてペアの回路で送信されるので、この設計では回路ペアの障害も検出することができます。SDRAM が故障すると、影響を受けたメモリ クワドラントは除外モードになります。除外モードでは、故障した SDRAM 以外のエラーも訂正対象になります。すなわち、別の SDRAM の故障も訂正対象になります。

このチップセットは、メモリの 1 ビット エラーに対する、ハードウェア ベースの自動スクラビング機能も備えています。この機能が採用されたのは、ソフトウェア ベースのスクラビングは本質的に信頼できないことが理由です。その一例として、ソフトウェア スクラバからは特定のアプリケーションでロック ダウンされたメモリにアクセスできない例があります。

ダブル チップ スペアリング (メモリ エラー訂正)

HP システムで改良された機能が、ダブル チップ スペアリング テクノロジーです。特定の DRAM 内のエラーは、ハードウェアで常に検出/訂正が可能です。また、同一コード ワード内の他のロケーションにあるメモリ エラーも検出/訂正が可能です。これはファームウェアによって実現しています。最初の DRAM が故障したとき、そのビットは ECC 訂正計算の対象から外されます。これにより、ECC ロジックは、同一 ECC コード ワード内の別の DRAM 障害を訂正できるようになります。この機能は、1 つの DRAM、バスの 1 ビットを共有しているすべての DRAM、またはメモリ サブシステムのすべてのバスに適用されます。したがって、この独自の保護メカニズムは最大限適用されることになります。この改良機能は、HP のメモリ テクノロジー分野の高度な研究の成果であり、HP sx1000 チップセットの単一チップ スペアリング サポートとは異なって、MB 単位の余分な DRAM を必要としません。この方式は、メモリ ミラーリング メソッドによるメモリ保護に比べるとはるかに費用対効果が高く、さらにメモリ ミラーリングが備えている利点をほとんどすべて備えています。この高可用性機能の主な利点は、故障した DIMM を交換するときのシステム ダウンタイムを削減できるということです。実際、このテクノロジーを使用すると、チップ スペアリング テクノロジーのみを使用しているシステムに比べて、DIMM の交換回数がおよそ 17 分の 1 に改善されます。

図 3. システム例: FRU (field replaceable unit) の修理回数



独自に DIMM を改良

HP の厳格な高可用性設計基準に従えば、メモリ アドレス、メモリ制御信号 (たとえば、RAS/CAS/WE)、およびタグ コマンドのパリティ エラーは、HP sx2000 チップセットで検出することになります。パリティ保護機能がないと、アドレス ラインや制御ラインで 1 ビットが反転したときに*正しい*データがメモリの*誤った*場所に書き込まれ、その結果データが破壊されることがあるので、このチェックは重要です。しかも、この種のデータ破壊は、誤ったデータが書き込まれた特定のメモリ ロケーションを調べる方法がないため、検出が非常に困難です (アプリケーション エラーだったのか？ ディスク エラーだったのか？ または、I/O エラーだったのか？ など)。 最終的に、HP の独自 DIMM には、重要なアドレス信号や制御信号に対して、冗長で表面積の大きい接点を取り付けました。これによってパリティ エラーの発生率が大幅に減少しました。これは業界標準の DIMM にはない機能です。

メモリの動的割り当て解除 (Dynamic Memory Resiliency: DMR)

コンピュータ システムのライフ サイクルにおいて、DRAM で 1 ビット故障 (永続故障) が発生することは、よくあります。永続性の 1 ビットのメモリ エラーが性能問題の原因にならないように、故障したビットのロケーションをメイン メモリのマッピングから除外します。

HP は、このシングル ビットの物理障害の問題に対処するため、HP の高可用性設計基準に従って独自のソリューションを開発しました。すなわち、メモリ内の特定のロケーションに故障があれば、その物理ページ (4k チャンク) を割り当て対象から外します。この物理ページは、リブートしても、割り当て対象から除外されたままです。システムのライフ サイクルにわたって、多くのメモリ ロケーションで永続故障が発生しますが、このような故障のほとんどは暗黙裡に処理されます。このような割り当て解除は動的に実行され、OS またはアプリケーションの介入を必要としません。メモリ内の特定ロケーションに問題がある (すなわち、永続故障を示す) ことが分かれば、そのメモリはオンラインのまま割り当てが解除され、ユーザには何の影響もありません。システム内の予備ページの数によっては、製品のライフ サイクルにわたって、故障したメモリを交換する必要がなくなるので、計画的なダウンタイムを大幅に減少させることができます。

この DPD 機能 (Dynamic Page De-allocation: ページの動的割り当て解除) は、本セクションで説明している他の機能と合わせて、DMR (Dynamic Memory Resiliency: メモリの動的割り当て解除) と称しています。

I/O サブシステムの RAS 機能

I/O アーキテクチャの概要

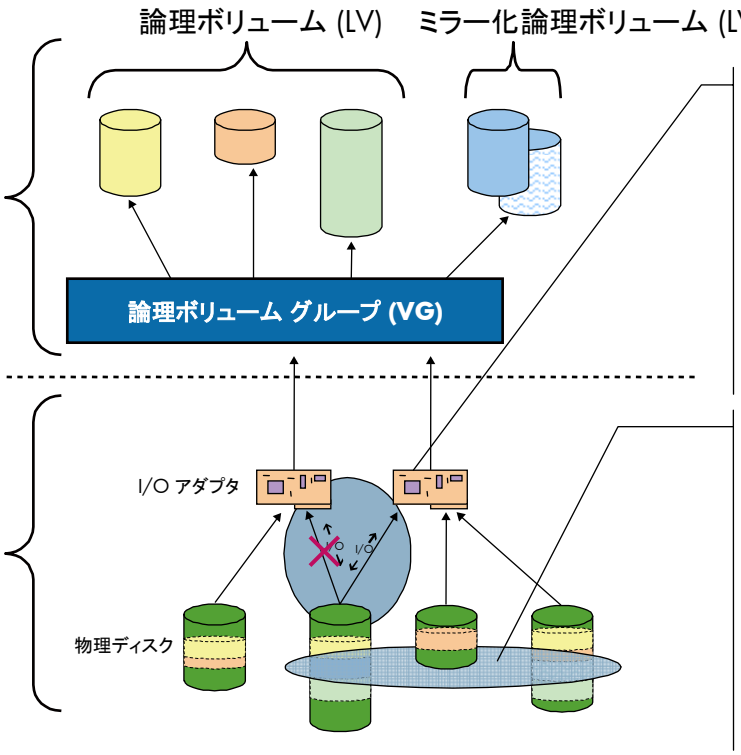
I/O サブシステムにより、プロセッサとメモリが I/O カードに接続されます。I/O サブシステムは、PCI-X システム バス アダプタと PCI-X ホスト ブリッジの 2 種類のチップで構成されています。セル コントローラは、高速リンクを経由して、1 つの PCI-X システム バス アダプタに接続されます。このリンクには、クロスバー ファブリックと同じ自己修復機能が装備されています。これにより、コネクタやケーブルの大部分の障害は、性能を著しく低下させることなく、修復することができます。

PCI-X システム バス アダプタには、8 ～ 12 個 (プラットフォームに依存) の PCI-X ホスト ブリッジが接続され、各 PCI-X ホスト ブリッジには、1 つの PCI/PCI-X バスが接続されます。この場合、共有スロットは使われません。このアーキテクチャには大きなシステム上の優れた可用性があります。これにより 1 つの PCI カード上で発生する障害は、1 つの PCI バスおよび PCI-X ホスト ブリッジから分離できます。I/O カードは、論理的にも電気的にも他のすべての PCI カードから分離されているので、PCI エラーは通常、エラーが発生したカードのみに封じ込められるため、エラーがシステム内の他の I/O カードに伝播することはありません。このため、I/O エラーがシステム全体の問題になる可能性は大幅に減少します。さらに、別のハード パーティションに伝播するような I/O カード エラーもありません。HP sx2000 チップセット内の PCI-X ホスト ブリッジでは、前バージョンの PCI/PCI-X の他に、PCI-X 2.0 266MHz もサポートしています。新しい PCI-X 2.0 I/O バスでは、広帯域幅になっただけでなく、ECC も追加されたので、信頼性が向上しています。HP sx2000 チップセットを使用すると、ECC をサポートするすべての PCI-X カードで、133MHz または 266MHz のいずれで動作する場合でも、ECC が有効になります。

マルチパス I/O

この機能を使用すると、複数のパスを経由してストレージ デバイスやネットワーキング エンド ノードにアクセスできるようになります。これは、同時アクセス (アクティブ/アクティブ構成の場合)、または簡易アクセス (アクティブ/パッシブ構成の場合) で行われます。この機能によって、2 つのエンド ポイント間の障害は排除されます。システム ソフトウェアは、ネットワークやストレージのリンク障害を自動的に検出してスタンバイ リンクにオンラインでフェイルオーバーすることができます。また、すべての I/O エラーの 90% 以上を占める I/O ケーブルやデバイス側の I/O カードのエラーに対して、システムの耐障害性機能が実現されています。

図 4. マルチパス I/O



PV リンク

同一デバイスへのマルチパス、1 つのパスはスタンバイ

- I/O 障害時には、LVM が最適の代替リンクを検索
- リンクの優先度は指定可能、LVM は優先リンクが復旧すると自動的に切り替える
- クラスタでは、アクティブ/パッシブ リンクの切り替えは当該デバイスを使用中のすべてのクライアント間で調整される

ミラー化論理ボリューム

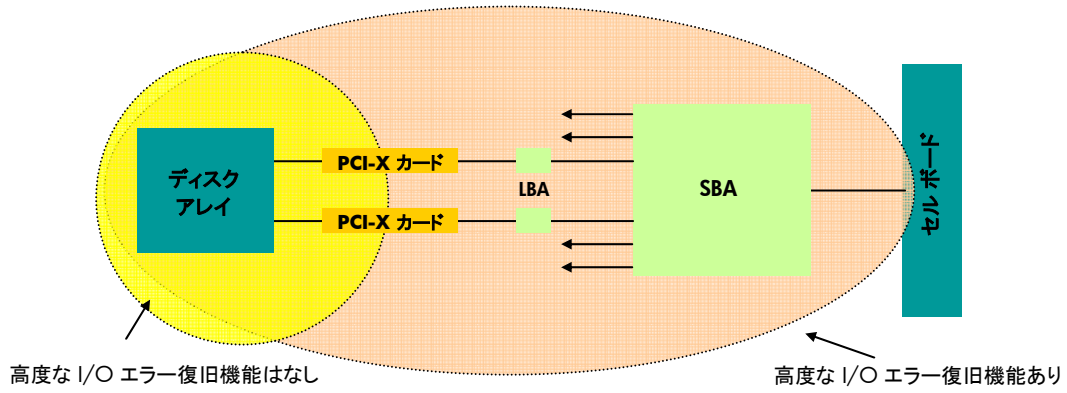
- ミラーリングは領域単位であり、ディスク単位ではない
- 3 つのミラーリング整合性ポリシー: MWC、完全再同期、なし
- 3 つの割り当てポリシー: 厳密、PVG 厳密、非厳密
- ミラーは分割や併合が可能
- すべてのコピーは同一 (優先度や順序はなし)
- 1 つのコピーしかできないときでも、読み書きは可能

高度な I/O エラー復旧

PCI エラー (たとえば、パリティ エラー) が発生した場合、PCI エラー処理機能により、HP-UX 11i システムのマシン チェック アボート (MCA) または高優先度マシン チェック (HPMC) を回避できます。PCI エラー処理機能を実装していない場合には、PCI スロットにはハード フェイル モードが設定されます。スロットがハード フェイル モードのときに PCI エラーが発生すると、MCA または HPMC が発生し、パーティションはクラッシュします。

PCI 高度エラー処理機能を実装している場合には、PCI エラー処理機能をサポートしている I/O カードが取り付けられている PCI スロットには、ソフト フェイル モードが設定されます。スロットがソフト フェイル モードのときに PCI エラーが発生すると、そのスロットはその後取り付けられた I/O カードから分離され、対応するデバイス ドライバがエラーを報告してから、サスペンド状態になります。スロット、カード、およびドライバを使用可能な状態に戻すには、olrad コマンドとアテンション ボタンを使用してオンラインで復旧させます。PCI 高度エラー処理機能とマルチパス機能を組み合せることにより、システム ダウンの 90% 以上の I/O エラー原因を排除できます。

図 5. 高度な I/O エラー復旧



PCI OL*

システム ハードウェアでは、オペレーティング システムがサポートする PCI カードの OLA 機能を組み合せることにより、スロット単位の電源制御が行われます。この機能を使用すると、他のコンポーネントに影響を与えることなく、またリブートの必要もなく新しいカードを追加することができます。また、I/O アダプタを追加するときにも、システムをアクティブ状態にしておくことができるので、お客様に対する高可用性ソリューション全体が強化されます。すべての PCI カード (Gigabit Ethernet、ファイバ チャネル、SCSI、Term I/O など) と、対応するドライバがこの機能をサポートしています。新しく追加したカードは、オンラインで構成してアプリケーションで使用することができます。PCI OL* は、HP 製品で簡単に使用できる機能であり、ドアベルやラッチを追加することで改良が行われています。これは、ときとして、ブラインド プラグ手順を使わざるを得ない業界標準の PCI サービス機能との際立った違いです。

また、時間が経過することにより I/O カードが故障した場合には、自動的に 2 番目のパスへのフェイルオーバーが行われたり、重要ではないデバイス (デュアル パス I/O を保証していないデバイス) との接続が解除されます。PCI オンライン交換 (OLR) 機能により、お客様はダウンタイムを認識することなく、故障した I/O カードを修理してオンラインでシステムを初期状態に戻すことができます。

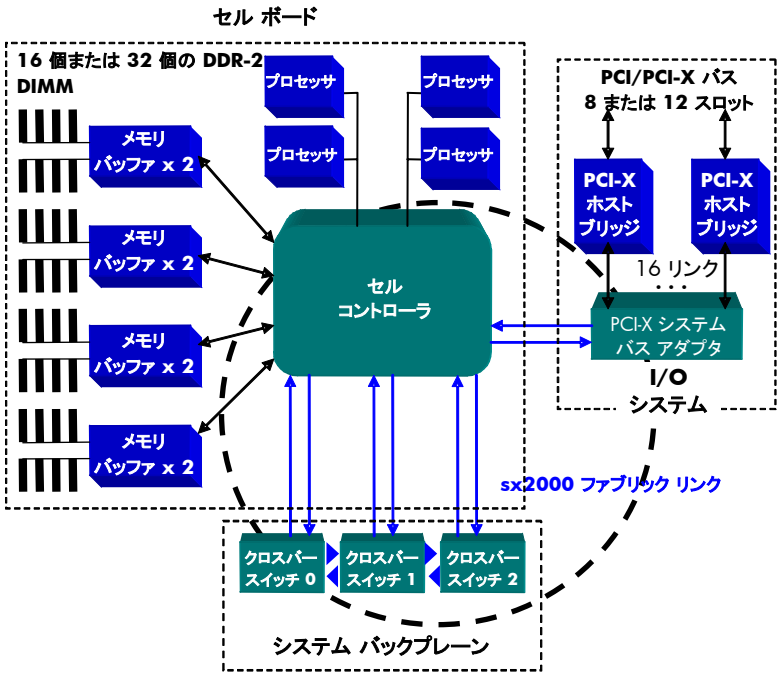
システム通信リンクの RAS 機能

本セクションでは、いくつかのチップ間通信機能について説明します。これらの機能が、HP の堅牢で高度なシステム ファブリックを実現しています。

リンク レベルのリトライとリンクの再学習

HP のチップセットは、リンク エラーを検出するとエラー イベントを通知し、リンクの再学習を行わせてから再試験を行い、通信パケットのうち、正常に送信されたパケットの確認を行って、その後通常動作に戻ります。これらの復旧イベントは、システムの動作に影響を与えることなく処理されます。エラーが検出された場合には、エラーを受信したチップが現在のパケットに対して無効情報を通知します。さらに後続するパケットもすべて無効として扱われます。十分なクロック サイクルが経過すれば、エラーはリンクの両端で認識されます。ここでリンクは、ブート時の電源投入時の初期化で使用されるのと同じシーケンスを使用して、再学習と再試験を開始します。これによってリンクは、前回の学習シーケンス以降に発生したドリフト状況や障害状況に合わせて調整されます。リンクの自己試験を使用して、障害が単一の一時的な障害か、永続的な障害であるかを判断します。学習シーケンスと自己試験シーケンスにより動作モードのリンクを適正に設定することができない場合には、致命的なエラーが発生し、リンクではダウン状態が継続します。ほとんどの場合、リンクは正常に再設定されます。再設定が完了すると、2 つのチップは、最後の正常トランザクションから通信を再開 (リンク リトライ) します。

図 6. リンク レベルのリトライ



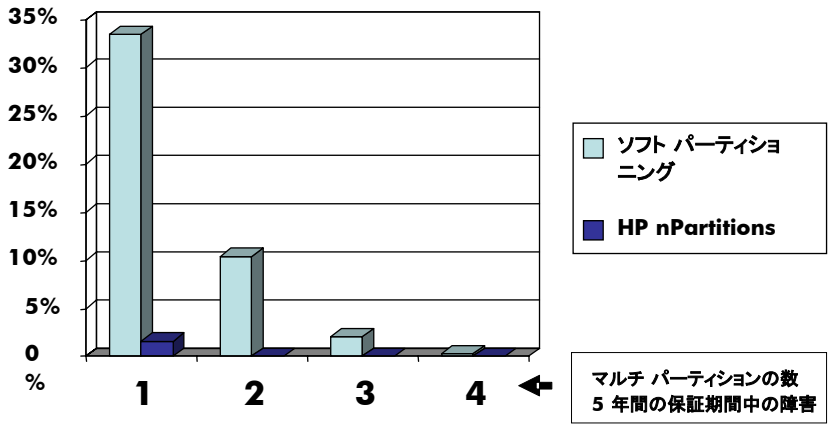
予備のチャネル/予備の回路

システムの各通信リンクは、20 個の並列チャネルで構成されていますが、通常、データとパリティの送信に必要なチャネル数は 19 個のみです。20 番目は予備ですが、通常動作においては、付加パリティ情報を送信するために使用されます。リンク学習 (または、再学習) 時にチャネルが未接続であると判定されると、予備チャネルが使用されて、不良チャネルは未使用のままとなります。これにより各リンクは、1 つのチャネルに障害があっても機能することができます。

キャビネット機構

キャビネット機構は、電源を投入すると、システムのライフ サイクル全体が終了するまで電源投入状態が継続することを想定して設計されています (すなわち、単一点障害の最小化)。

図 7. 複数のパーティションが障害を起こす確率

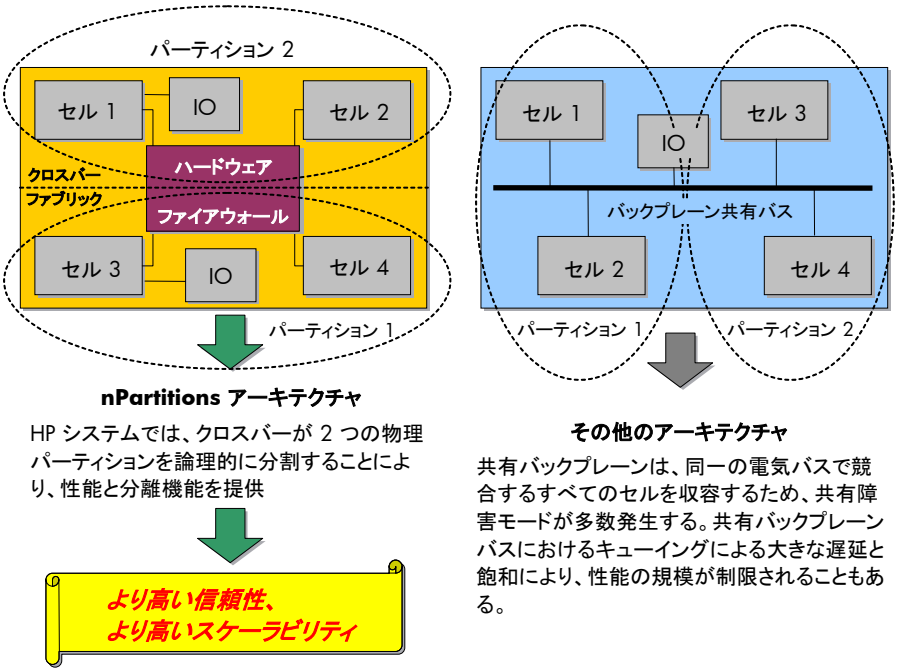


全負荷状態でパーティション化されたプロセッサ 32 個構成のサーバの HP nPartitions は、5 年間の保障期間中で、ソフト パーティションに比べて **40 倍以上の可用性**を提供します。

クロスバー バックプレーンの障害復旧力

システム クロスバーは、単一パーティション システムの高信頼性と共に、パーティション間における前例のない封じ込め機能を装備しています。これは、クロスバー チップセットにハイグレードな部品を使用し、バックプレーンにはホットスワップ機能と冗長電源を備え、HP Integrity Superdome のセル、バックプレーン、および I/O との間には準耐障害通信パスを使用することで実現しています。また、HP では、パーティション機能を持つ他のシステムとは異なり、故障したパーティションが生成する不正なトランザクションからパーティションを保護するために、独自の専用ハードウェアを提供しています。

図 8. ハード パーティションのエラー封じ込め機能



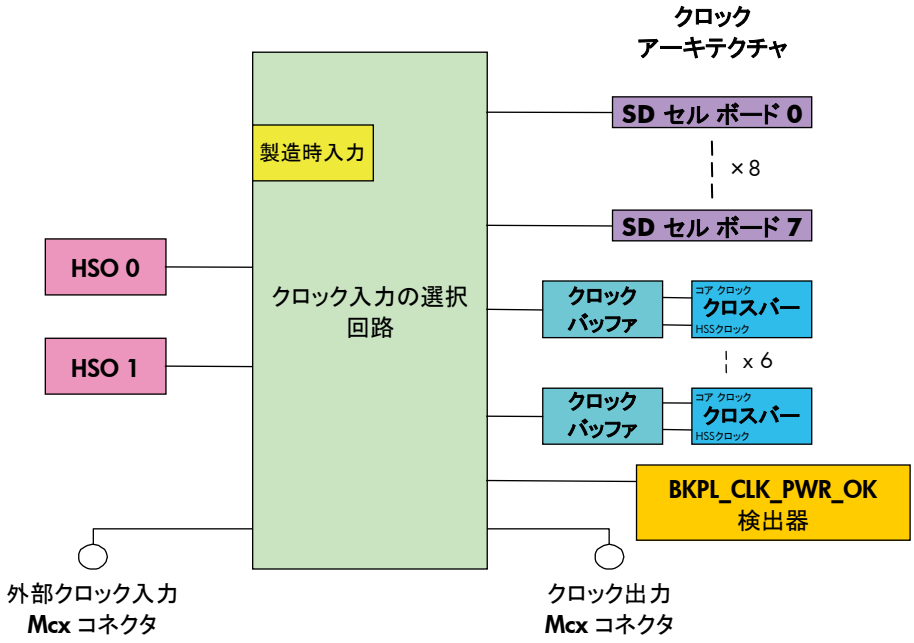
このような復旧機能は、真の意味でのハード パーティションに必要です。各ハード パーティションには、パーティションを形成するセルのリソースで構成される、専用の独立したプロセッサ、メモリ、および I/O リソースがあります。リソースは、システム管理インタフェースに含まれているコマンドを使用して、ハードウェアを物理的に操作しないで 1 つのパーティションから削除して他のパーティションに追加することができます。HP-UX 11i の今後のリリースでは、動的再構成 (たとえば、オンライン追加、オンライン削除) の関連機能を使用して、新しいリソースをパーティションに追加したり、故障したモジュールはパーティションを動作させたまま交換できるようになります。

クロックの冗長化

システム クロックは、完全に冗長なホットプラグ対応の 2 個のハードウェア基準オシレータ (HSO) から生成します。HSO は、自動で誤作動のないフェイルオーバー/再構成機能をサポートしており、システムのいかなる動作状況においてもホットプラグ対応です。

通常の操作では、システムは 2 つの HSO のうちの 1 つをプラットフォーム用のクロック ソースとして選択します。オシレータがバックプレーンに取り付けられていて、出力レベルが適正な HSO が選択対象になります。取り付けられている HSO が 1 つで、その出力が適正な振幅であれば、それが選択されます。HSO の出力が適正な場合には、HSO の緑色の LED が点灯します。HSO の出力が適正でない場合には HSO の黄色の LED が点灯し、アラーム信号がシステムに送信されます。両方の HSO が取り付けられていて、いずれの出力も適正な振幅である場合には、RCS のロジックに従って、2 つのうちの 1 つがクロック ソースとして選択されます。この場合、両方の HSO の緑色の LED が点灯します。

図 9. クロックの冗長化

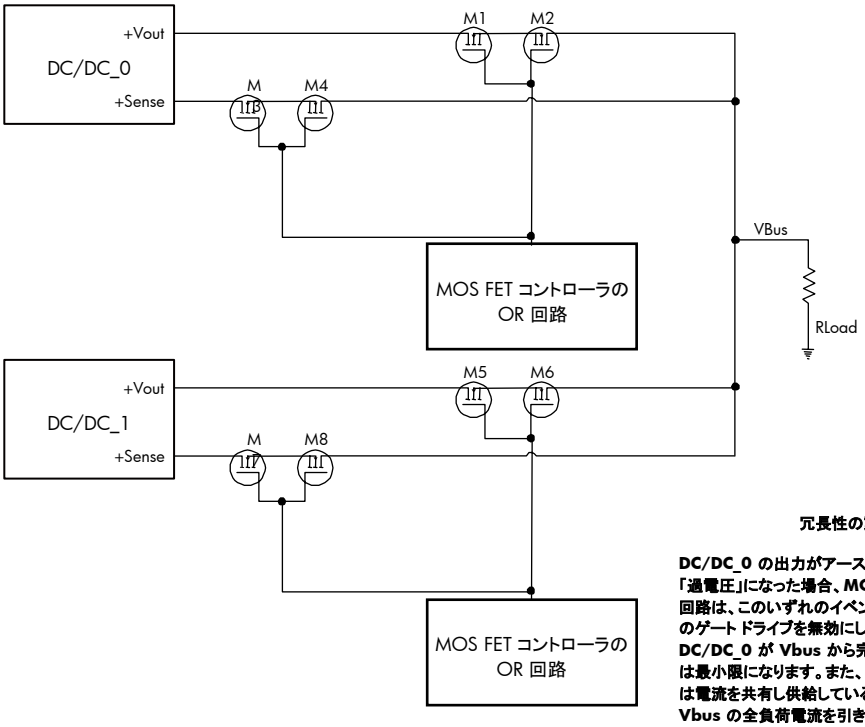


一方の HSO の出力が適正な振幅でない場合には、RCS は正常な HSO をクロック ソースとして使用し、故障している HSO を示す警告信号をシステムに送信します。正常な HSO では緑色の LED が点灯し、故障している HSO では黄色の LED が点灯します。故障しているクロック ソースはホット プラグ操作によって修復することができます。

電源機構

HP sx2000 チップセット搭載システムの大部分の DC-DC コンバータには完全な冗長機能があり、必要な部品点数を削減することにより従来はパレート図における重要な項目の 1 つであった電源変換機能に伴うダウンタイムを大幅に減少させています。

図 10. 電源機構



冗長性の動作説明

DC/DC_0 の出力がアースにショートするか、または「過電圧」になった場合、MOS FET コントローラの OR 回路は、このいずれのイベントも検出し、M1 ～ M4 へのゲート ドライブを無効にします。これにより、DC/DC_0 が Vbus から完全に分離されるので、影響は最小限になります。また、DC/DC_0 と DC/DC_1 は電流を共有し供給しているので、DC/DC_1 が Vbus の全負荷電流を引き継ぐことになります。

また、HP sx2000 チップセット搭載の HP Integrity サーバのハードウェアは、AC 入力を 2 台の異なる AC 電源から供給することができます。この目的は、電源 A と電源 B の両方、電源 A のみ、または電源 B のみのいずれの給電パターンでも、すべての装置を正常に動作継続させることです。この機能は「AC 入力電源に関する耐障害性の規定基準」と呼ばれ、Uptime Institute からの認証を取得しています。

装置が Uptime Institute からこの基準に適合していると認定されるには、最初の設置段階から最終的な運用段階まで、次のすべての基準を満たす必要があります。

- 2 台の AC 電源のいずれか一方が故障した場合、または耐障害性の基準を逸脱した場合でも装置を起動することができること、またはデータを損失したり、ハードウェアの機能、性能、キャパシティ、冷却機能を損なうことなく、中断なしで動作を継続できること。
- 連続使用している電源が他の AC 電源から供給されているときに、一方の AC 電源が故障している状態、または耐障害性を逸脱している状態から回復した場合に、電源を切断したり、IPL を行ったり、あるいは人手の介入なしでデータ、ハードウェアの機能、性能、キャパシティを復旧できること。
- 1 番目または 2 番目の AC 電源が故障している状態、または耐障害性を逸脱している状態から回復したことに伴って一時的に中断しても、データを損失したり、ハードウェアの機能、パフォーマンス、キャパシティ、冷却機能を損なうことがないこと。
- 2 台の AC 電源で互いに同期をとる必要がないこと。すなわち、各 AC 電源の電源特性が製造メーカーが公表している仕様と許容値の範囲である限り、電圧、周波数、相回転、位相角が異なっていても構わない。
- 両方の AC 電源入力は、製造メーカーの電源装置内で終端してあること。内部または外部の能動入力切替装置 (たとえば、静的切替スイッチ) は、使用できません。

- 製造メーカーの電源装置内で障害により一方の AC 電源が故障しても、その故障が他の AC 電源に波及して故障させる原因にならないこと。
- 両方の AC 電源入力を使用している場合、2 つの内部パワー トレインから供給される電源は、電源出力の 50% ± 10% であること。
- AC 電源が故障したり、製造メーカーが公表している許容値を逸脱したとき、および異常状態が修復されたときに、装置のソフトウェアまたはホストのオペレーティング システムを通じて、外部のソフトウェアに警告を送信する機能が提供されていること。

HP Integrity システムの個々の BPS (Bulk Power Supply: バルク電源) の設計と電源の相互接続構成は、上記の「AC 入力電源に関する耐障害性の規定基準」を満たしています。

冷却機構

システム内のすべてのファン (すなわち、システム ブロワと IO カード ケージ用のファン) には、完全な冗長機能がありホットスワップ対応です。これらの部品は、システム内のどのパーティションにも影響を与えることなく、保守できます。

管理プロセッサ

MP (Management Processor: 管理プロセッサ) は、保守用およびコンソールとして使用できるインタフェースの 1 つです。MP は、システムのヘルス状況を監視する方法の 1 つでもあります。MP の明確な利点の 1 つは動作環境に依存しないで使用できることであり、オペレーティング システムとペイロード LAN の両方に依存しないで使用できます。MP は、「Extended Fault Management (広範囲をカバーする障害管理機能)」と呼ばれることがあります。これは、リモート電源制御や、キャビネット、セル、I/O シャーシのファン、温度、電源装置などの環境監視、システム構成の表示、MP の構成制御などの機能を持っているためです。ユーザは MP をパーティション コンソールに接続して、サーバやパーティションの状態を表示したり、パーティションをリセットしたり、アーカイブされているログや現在のログを表示したり、その他各種のプラットフォーム管理機能を実行することができます (40 個以上のコマンドが用意されています)。MP は、ユーザを認証し、ユーザに役割ベースの機能を割り当て、セキュアな組み込み Web コンソールを使用するセキュアな (SSL ベース) アクセスを提供することにより、権限のないユーザがシステムのこれらの機能を使用できないようにします。

MP には VFP (Virtual Front Panel: 仮想フロント パネル) も用意されています。これは、各パーティションのブート ステータスと動作をリアルタイムで表示します。VFP は多くの単一パーティション サーバにある LCD ディスプレイに類似していて、セルまたはパーティションのブート シーケンスの進行状況、各セルまたはパーティションの動作状況、およびブート プロセスで検出されたエラー、フラグ付けされたエラーなどの情報を表示します。

MP はシステムの可用性を監視したり、制御するように設計されています。つまり、MP は、MP が故障しても、パーティションで発生するクラッシュを防止したり、パーティションでブートやリブートを行うことができるように設計されています。また、MP はオンラインで交換できるので、MP の故障によるダウンタイムは発生しません。他社にも MP を備えているシステムはありますが、MP の保守時にはシステムをダウンさせる必要があります。

グローバルな共有メモリ

GSM (Global Shared Memory: グローバルな共有メモリ) を使用することにより、2 つの異なるパーティション間で、システム ファブリックを経由して、瞬時に交信することができます。HP sx2000 チップセットの GSM は、複数の同種オペレーティング システム間 (Windows と Windows 間)、または異種オペレーティング システム間 (Windows と HP-UX 間) の同時通信をサポートしています。この機能は、複数のサーバを 1 つの HP sx2000 チップセット搭載システム内の複数のハード パーティションに集約する場合に、特に役に立ちます。各パーティションでは、スタンドアロン サーバで実行されたアプリケーションと同じアプリケーションを続行することができます。また、パーティション間の通信は、ネットワーク カード (NIC) を使わずに、低遅延、高帯域幅のシステム ファブリックを経由するため高速です。

HP sx2000 チップセットそのものに設計されているセキュリティ機能を使用してハード パーティション間にファイアウォールが保持されることは重要です。これにより、1 つのパーティション内のアプリケーション (または、故障) によって他のパーティションが停止することはありません。ホスト パーティションとは、GSM メモリを所有しているパーティションのことです。ゲスト パーティションとは、ホスト パーティションが提供するリモート メモリを読み込むパーティションのことです。ゲスト パーティションはリモート メモリを読み込むことはできますが、書き込むことはできません。リモート パーティション内のハードウェア障害やソフトウェア障害は、メモリが共有されている場合でも分離されます。GSM が使えるかどうかは、オペレーティング システムに依存します。

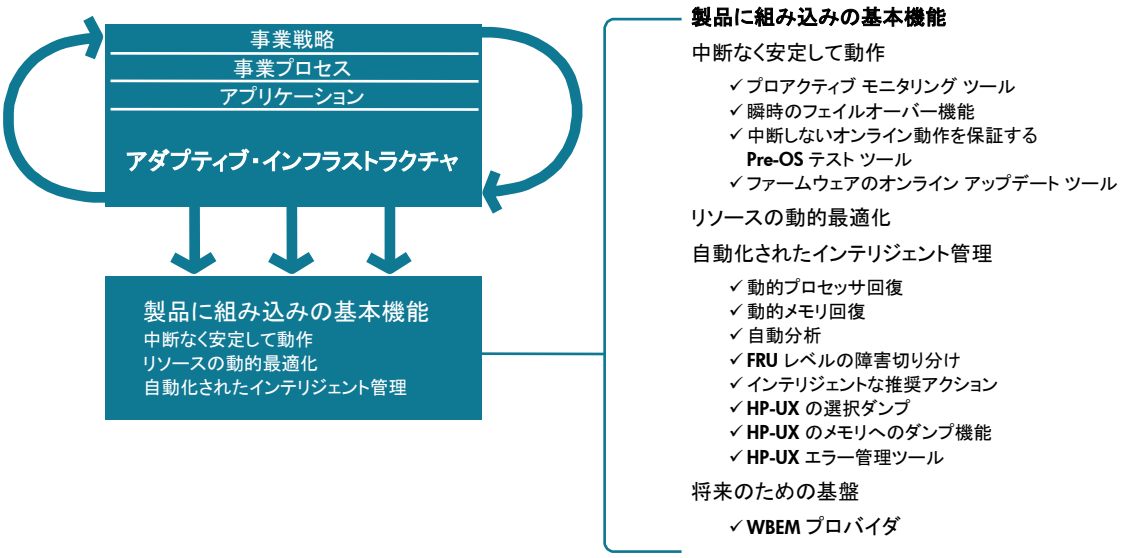
ファームウェアの更新

各ハード パーティションにはシャドウ ROM が含まれており、パーティションをオンラインの状態にしたまま、ファームウェアの更新を実行させることができます。新しいファームウェアは、次回のパーティション レベルでのリブート時に実行されます。ファームウェアのバージョンはハード パーティション間で一様である必要はありません。そこで、たとえば test という名前のパーティションのファームウェアを最初に更新し、その後で残りのパーティションのファームウェアを必要に応じて更新するといった、段階的なアップグレードが実行可能です。「キャビネットへの電源投入は 1 回のみ (power-on cabinet once)」という設計方針に従い、キャビネット レベルのリソース (MP、Event Dictionary、キャビネット レベルのユーティリティ) はいずれも、OS の停止やリブートを要することなくオンラインの状態でファームウェアを更新できるように設計されています。

障害管理

障害管理とは、システム可用性が予想に反して損なわれたことを検出して対策するアクション（*予防型*、*リアクティブ型*、*プロアクティブ型*、*瞬時型*）のことです。前述したように、障害管理はすべての HP サーバの品質、設計、およびエンジニアリングを検討することから始まります。

図 11. 障害管理 — 製品の基本機能に組み込まれたアダプティブ・インフラストラクチャ



適応型システムの開発では、製品に組み込む基本機能は、障害管理を念頭において設計されます。こうすることにより、中断しないセキュアな運用、リソースの動的最適化、自動化されたインテリジェント管理を実現することができます。これが、自己修復とプラットフォーム管理を可能にするソリューションの基本です。これは、新しい考え方ではありませんし、新たなイニシアティブでもありませんが、文字どおり常に HP のサーバ設計の基本です。インテルの Itanium プロセッサ ファミリーが出荷されたことによって、HP ではこの基本的なデザインを拡張して、Windows や Linux にも適用できるようにしました。その結果提供されることになった障害管理機能は、以下のような 4 つの顧客価値として言い換えることができます。

<u>障害が原因のダウンタイムのリスク低減</u>

- ハードウェア検証と潜在的な部品故障を洗い出すためのバーンイン ツール
- 自己修復機能と高信頼性のハードウェア

<u>障害の修復に要する最小ダウンタイム</u>

- 直観的なツール
- 問題点の検出と切り分けに対する比類のない適用範囲の広さと深さ
- 報告イベントには問題解決に対する明確で正確な推奨アクションが含まれているため、ハードコピーのガイドを参照することは不要
- 間欠的なパフォーマンス低下問題に対するテスト ツール
- 修理が適正に行われたことを確認する検証ツール
- 物理的な修理を簡単に短時間で行えること

<u>問題解決が簡単で低コストでできること、すなわち TCO の低減</u>

- 簡単にインストールできるように、ツールがインストール済みであること

- HP の OpenView や IBM の Tivoli などのエンタープライズ管理アプリケーションが使用されている場合に、それらの既存のインフラストラクチャに簡単に統合できること
- HP でお客様の環境を完全に管理する場合には、障害管理ツールは、ISEE (Instant Support Enterprise Edition: インスタント サポート エンタープライズ エディション) の機能を拡張したサポート サービスを行うための欠かせないツールとなります。
- 障害管理アプリケーションは、直観的で効果的な操作を行えるように設計されています。ユーザ フレンドリなイベントの説明文により、問題判定ツールを簡単に使用できて理解することができます。
- MP はリモート障害管理の機能も備えているので、従来はなかった方法で HP サーバのトラブルシューティングを行ったり、管理することができます。
- 保守性: ハードウェアの交換や修理手順の削減
- 耐障害性や障害封じ込め設計により、修理が簡単です。

操作性の向上

- お客様の使い勝手が、以下の機能により改善されています。
  - 単純さ
  - HP Integrity Superdome のハードウェアで使用するオペレーティング システムには依存しない一貫性のある使い勝手

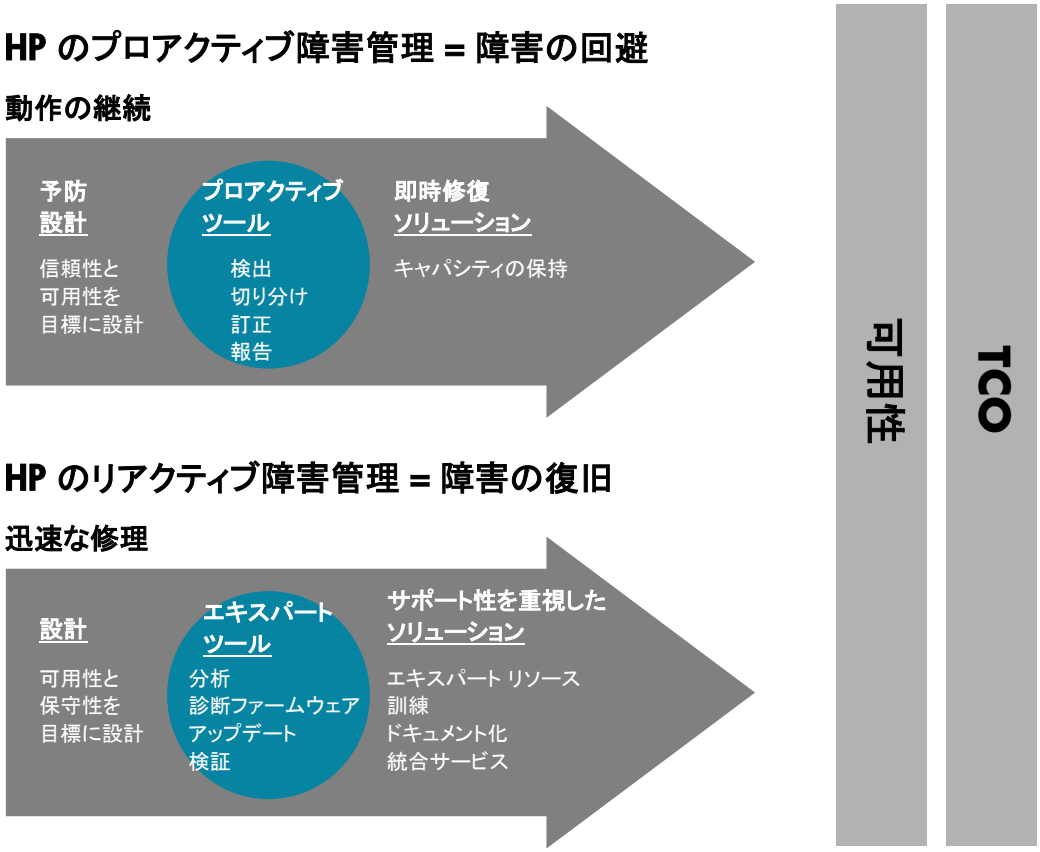
保守性

保守が部品を交換することだけを指していたのは過去のことです。今では、保守は問題に対する効果的な対策、復旧、および解決を実現するためのソリューション全体を指すようになっています。保守性は、他の RAS 要素と同様に、ツール、プロセス、ハードウェア設計で構成され、最終的にはシステムの可用性に寄与し TCO の決定要因になります。

HP の保守性についての戦略は単純です。すなわち、迅速に修理できることです。問題解決は、より迅速に、より簡単に、より正確に行われます。HP は、障害が発生したとき、あるいは障害によってハードウェアを交換しなくてはならない場合にも、最大限の稼働時間を確保できるソリューションを開発すべく、最善を尽くして取り組んでいます。

次の図は、障害管理を分類した中のリアクティブな復旧方法も示していますが、「動作を継続させる」ための予防型の側面も考察し、以降でプロアクティブ ツールについて詳しく説明します。

図 12. HP のプロアクティブとリアクティブ障害管理



ハードウェアの予防保守設計

障害管理設計では、ハードウェア設計者と OS 設計者が共同作業を開始し、ハードウェアとソフトウェアの機能分担を取り決めて、製造レベルとエンドユーザの本番環境の両方を対象にしたシステム異常を、検出して切り分けるための測定ポイントを作成します。HP Integrity Superdome には、お客様の高可用性環境で最大限の稼働時間を保証する最新で先進の HP 障害管理機能があります。障害を回避する設計手法の大部分は、信頼性と可用性のセクションで説明しました。以下のセクションでは、HP Integrity Superdome の保守性と障害復旧について、今までに説明しなかった機能を説明します。

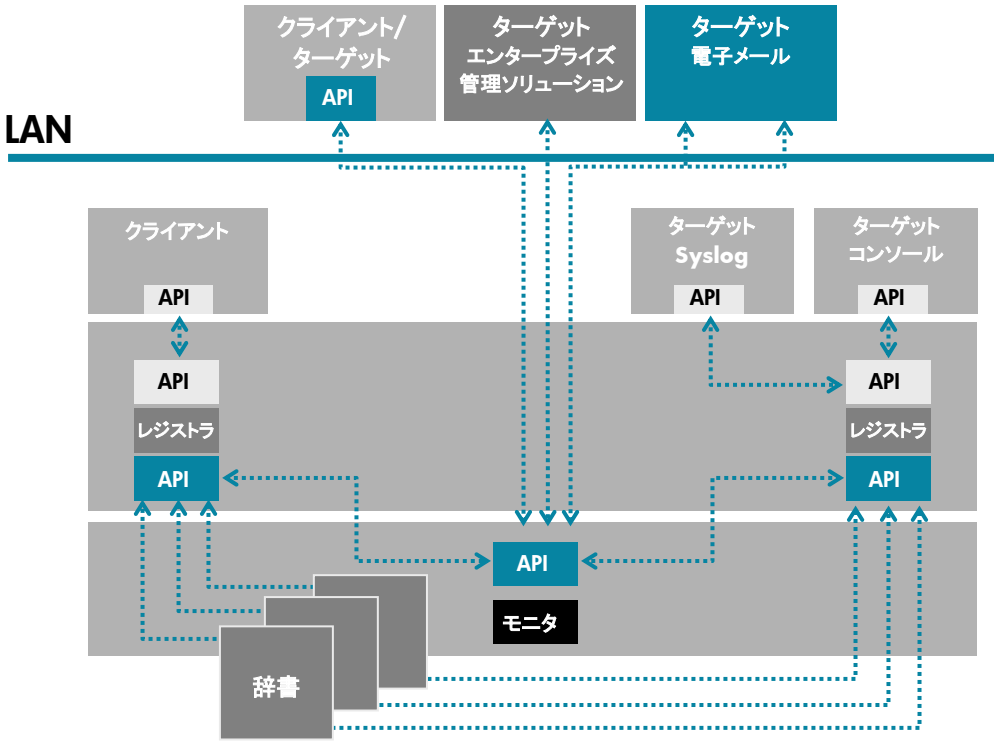
HP-UX のプロアクティブ ツール

プロアクティブ障害管理機能の大部分は、予防型のハードウェア設計に適切に組み込まれていますが、後続の障害を監視、検出、報告するためのツールもあります。監視ツールには、計画外のダウンタイム イベントが発生する前に、修復作業を行うための機能も含まれています。これらの監視ツールは、HP-UX EMS に含まれています。

HP-UX EMS

EMS (Event Monitoring Service) は、ターゲット アプリケーションとクライアント アプリケーション、リソース モニタ、および EMS フレームワークの 3 つの主要なコンポーネントで構成されています。クライアント アプリケーションは、モニタを構成するために使用するアプリケーションです。ターゲット アプリケーションは、通知の送信先であるエンドユーザ インタフェースのことです。EMS フレームワークには、API、レジストラ、リソース辞書があります。

図 13. EMS のコンポーネント



リソース モニタ

リソース モニタは、ハードウェア モニタ、ソフトウェア モニタ、およびオプションの高可用性モニタで構成されています。これらのモニタは、エージェントまたはプロバイダと呼ばれることもあります。リソース モニタは、システム障害または近々に発生予測される障害の検出、切り分け、報告を行います。通知は、システム コンソール、電子メール、ポケベル、ログ (テキスト ログまたは syslog)、HP サポート サービス、およびエンタープライズ管理プラットフォーム宛てに送信されます。

ハードウェア モニタはベース オペレーティング システムに組み込まれ、コア システム コンポーネントと周辺デバイスの両方をカバーします。ネイティブの EMS モニタがありますが、これには、プロセッサ、メモリ、シャーシ、ファン、システム バス、温度、ディスク アレイ、SCSI デバイスとエンクロージャ、Fibre Channel デバイス、および電源などのモニタが含まれます。モニタは、サブシステムのログ追跡、最も疑わしい FRU とのイベント相関処理、近い将来に障害が発生しないかどうかを調べる予測分析、障害の原因と対策についての報告を行います。また、モニタには、業界用語で自己修復機能と呼ばれる、切り分けと修復するための機能を提供することによりその価値を高めています。この機能には、故障したプロセッサのオンラインでの永続的な割り当て解除 (DPR)、メモリ ページの割り当て解除、電源が UPS にフェイルオーバーしたときの正常な手順に沿ったシステムのシャットダウン、およびシステム ハング時のシステムの自動再起動などが含まれます。

ソフトウェア モニタと高可用性モニタには、さらに広範囲のモニタリング機能が含まれています。これらのモニタは、カーソル リソース、ネットワーク、LAN の集合とフェイルオーバー、データベース、クラスタなどを対象にします。

まとめ

対象	機能	ユーザの観点
メモリ システム	エラー検出/訂正 DIMM のアドレス パリティ保護 独自 DIMM の改良 ダブル チップ スペアリングによる動的メモリ回復 電源冗長化	DIMM の交換回数が従来のチップ スペアリングに比べて 17 分の 1 に減少 システム性能の面で妥協せず、ハードウェアの追加コストも必要としない高レベルの可用性 HP の DIMM の改良によりメモリ データ破壊のリスクがほとんど 0 に激減
プロセッサ	キャッシュ エラーの検出/訂正 自己修復キャッシュ コア ロジックのパリティ保護 動的プロセッサ回復 プロセッサ バス ECC	すべてのキャッシュ エラーと、大部分 (70%) のプロセッサ コア エラーを訂正できるので、x86 プロセッサに比べるとエラーの訂正範囲とデータ整合性が著しく向上。エンタープライズ カスタマにとってはエンタープライズ クラスの信頼性を確保
セル、I/O、ファブリック リンク	リンク レベルのリトライ リンクの予備チャネル	耐障害性リンクによって、パーティションはダウンせず修理も不要。この機能によって、現場のサーバ障害の共通的な原因である、環境上の問題と潜在している製造上の欠陥に起因するエラーを除去
クロスバー/システム ファブリック	セルとの冗長リンク ハード パーティション機能の系統立てられたサポート 電源冗長性 ファブリック内の三重冗長化	HP がパーティション戦術でリーダーシップを維持していることの主要な要因
I/O スロット	エラー検出/訂正 PCI 障害の単一スロットへの切り分け 拡張 I/O エラー復旧 マルチ パス機能 PCI カードの OLARD 電源冗長化	システム ダウンの主要な原因である I/O エラー (約 2 年間の MTBF) を、問題にならないレベル (約 50 年間の MTBF) にまで改善。オンライン修理機能によって、障害回避機能をさらに強化
チップセット	内部データ パスのエラー検出/訂正	HP の高度なチップセットによって競合他社のどこよりも優れた性能と可用性を実現
パーティション機能	nPartitions — パーティション間でのハードウェアとソフトウェアの分離 分離を損なわない GSM 通信 セル OLARD 冗長性のあるホットスワップ クロック ホットスワップ対応のサービス プロセッサ ボード Uptime Institute の認証を受けた AC 電源対策 保守容易性 – 複数のパーティションをダウンさせずにハードウェアの修復が可能	真の意味でのサーバ コンソリデーション戦略を実現。300 年以上と算出された MTBF と、HP の 2 世代にわたるハード パーティショニング技術により、お客様は、ハード パーティションに分割されたサーバを、より小さな筐体でアレイ製品のように扱うことができるだけなく、システム管理や TCO に悩まされることもありません。

参考情報

HP Integrity サーバ:
www.hp.com/jp/integrity (日本語)
http://www.hp.com/go/integrity (英語)

HP パーティショニング関連:
www.hp.com/jp/partitioning (日本語)
http://h71028.www7.hp.com/enterprise/cache/257389-0-0-0-121.aspx (英語)

EMS モニタの詳細:
www.docs.hp.com/ja/hpuxha.html (日本語)
http://www.docs.hp.com/hpux/ha/index.html (英語)

安全に関するご注意　ご使用の際は、商品に添付の取扱説明書をよくお読みの上、正しくお使いください。水、湿気、油煙等の多い場所に設置しないでください。火災、故障、感電などの原因となることがあります。

お問い合わせはカスタマ インフォメーション センタへ
03-6416-6512 月～金9:00～19:00 土10:00～18:00 (日、祝祭日、年末年始および5/1を除く)
HP Integrity に関する情報は http://www.hp.com/jp/integrity

Intel、Itaniumは、米国およびその他の国におけるIntel Corporationまたはその子会社の商標または登録商標です。
記載されている会社名および商品名は、各社の商標または登録商標です。
記載事項は2006年11月現在のものです。
本書に記載された内容は、予告なく変更されることがあります。
本書中の技術的あるいは校正上の誤り、省略に対して、
いかなる責任も負いかねますのでご了承ください。
本書は、『RAS features of sx2000 based platforms』(英語) をもとに加筆・修正して日本語で提供するものです。
© Copyright 2006 Hewlett-Packard Development Company,L.P.

日本ヒューレット・パッカード株式会社
〒102-0076 東京都千代田区五番町七番地

PDFHS06035-02